Collected Technical Notes on Weighting and Its Impact on the 2002 Census of Agriculture

Phillip S. Kott

This is a collection of five technical notes on weighting for nonresponse, coverage errors, and sampling in the 2002 Census of Agriculture. Also addressed are the techniques used for integerizing Census weights, smoothing sample-based undercoverage estimates, estimating mean-squared-errors, and calculating the fractions of published tabulations attributable to coverage adjustment (complicated by how "late adds" were treated in the Census process).

KEY WORDS: Nonresponse; Coverage-adjusted; Integerization; Sample; Smoothed; Late adds; Corrected; Attributed.

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030.

August 4, 2004

This is a collection of five technical notes written by Dr. Phillip S. Kott, Chief Research Statistician of the National Agricultural Statistics Service, in support of the 2002 Census of Agriculture. They are

A. Weighting in the 2002 Census of Agriculture
describes how nonresponse weights were computed and integerized. The note also discusses sampling weights and re-integerization (a routine to minimize changes from one integerization to the next). For a treatment of the calibration program used in determining coverage-adjusted weights before (re)integerization, see http://www.fcsm.gov/03papers/fetter_kott.pdf.

B. Smoothing State Farm Counts for the 2002 Census of Agriculture
describes how the number of farms in each state, a primary calibration target, was determined.

C. A Components-of-Variance-Model Estimator for State-Level NML Proportions
describes computing the number of NML farms (farms not on the Census Mailing List) within a category used as a calibration target. The note also discusses variance (and covariance) estimation for the fraction of NML farms within a category.

D. *Mean-Squared-Error Estimates for the 2002 Census of Agriculture*

E. Determining the Attributed and Corrected Fractions of Nonreponse and Undercoverage Adjustment
explains how farms that did not respond to either the screening survey or the Census itself were treated when measuring the extent of nonresponse and undercoverage on the Census of Agriculture.

For useful background information of the Census process, see http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf. This US-level document (there are state-level analogues) is hereafter referred to as *Appendix C*.

Weighting in the 2002 Census of Agriculture

The Weighting Steps for a Census Record

Each in-scope Census record begins with a weight of 1 (a record is in scope if it represents a unique farm operation). There are four additional weighting steps for Census records:

A. *Nonresponse weighting* – a nonresponse weight for record k is denoted by a_k.
B. *Integerization* – an integerized nonresponse weight is denoted by $a_k^{(I)}$.
C. *Coverage adjustment* – a coverage-adjusted weight is denoted by w_k.
D. *Re-integerization* – an integerized coverage-adjusted weight denoted by $w_k^{(I)}$.

We will discuss each in turn before moving on to sample reweighting.

Nonresponse weighting

Every potential operation on the Census list frame (Census mail list) begins with a weight of 1. Operations that are designed as *musts* based on list-frame information retain that weight after nonresponse adjustment. There are three kinds of musts. Operations with list-frame total-value-of-products (TVP) or land-on-farm above a state-dependent cutoff, *tagged* records (Agricultural-Resources-Management-Study selections and state-determined specials cases), and abnormals (mostly prison and research farms). The term “tagged records” as used here includes many records not actually tagged during census processing but later determined to be special cases for weighting purposes.

Potential operations sent the screening survey (the 2002 Farm Identification Survey) and responding in scope to that survey are considered on the Census list frame. Potential operations sent the screening survey and either not responding or responding out of scope are *not* considered on the Census list frame.

If a nonrespondent to the screening survey was sent a Census form and responds in scope, the record receives a nonresponse weight of 1. NASS's measure of undercoverage is affected when a screener nonrespondent responds in scope to the Census, but that will be discussed elsewhere (Section E). Late adds – both post-screener and post-first-Census-mailing – are treated like screener nonrespondents.

A single operator may have been sent more that one Census form. (S)he was required to fill out a separate form for each operation. When an operator with an operation for which he had *not* received a Census form reported the existence of such an operation on at least one of his received forms, the newly discovered potential operation (it may yet be proven out of scope) is an OpDom or OD add. Each OD add was sent a Census form or enumerated via telephone. If it responded, the OD add's nonresponse weight is set equal to the nonresponse weight of its *donor* – the operation reporting its existence (if there is more than one such operation, that is, if an operator fills out two or more Census forms reporting the existence of the missing operation, then a donor is selected at random from among them).

Some farms operating in more than one county are treated as several distinct operations. When the decision to create an additional record of this sort was made *after* the original operation responds to the Census, the *split add* receives the same nonresponse weight as the original record.

In-scope records that are *not* OD adds or split adds are nonresponse weighted using the following formula:

$$a_k = N_{NR(k)} / R_{NR(k)}, \qquad \text{(A1)}$$

where $N_{NR(k)}$ is the number of potential list operations in the same nonresponse group as operation k *not counting those undeliverable as addressed*, and $R_{NR(k)}$ is the number of potential list operations in the same nonresponse group as operation k that return Census forms (by "return a Census form" or "respond to the Census," we mean respond in such a way that NASS can either use the information as a valid record or determine that the respondent is a duplicate or a nonfarm; note that data for *musts* are imputed from secondary sources if need be). Both these counts *exclude* OD adds and *include* out of scopes – nonfarms and duplicates. Only a single record of an operation returning multiple forms is considered in scope, however, and receives a positive a_k value.

Nonresponse (NR) groups are formed within *list-frame counties*. An operation's list-frame county is the county NASS believes the operation to be in at the time it was mailed a Census form. We use list-frame rather than reporting county to form the NR groups because we do not know the reporting county of the nonrespondents.

Within a list-frame county, we have the following NR groups:

Group 0: *Must* records

Group 1: Expected 2002 TVP (based on list-frame information) less than $2,500

Group 2: Expected 2002 TVP between $2,500 and $9,999.

Group 3: Expected 2002 TVP between $10,000 and $49,000 and previously reported survey data from 1997 or later.

Group 4: Expected 2002 TVP greater than or equal to $50,000 and reported survey data from 1997 or later.

Group 5: Expected 2002 TVP greater than or equal to $10,000 and no reported survey data from 1997 or later

Any NR group within a list-frame county with less than two respondents (*including* duplicates or other out of scopes) or with less than half of the mailed-and-delivered-to operators responding is collapsed if possible. The collapsing pattern is:

Group 5 to Group 4 or 3 as appropriate

Group 4 to Group 3

Group 3 to Group 4 (call the result Group 3)

Group 3 to Group 5 (if Group 5 is not already collapsed; call the result Group 3)

Group 3 to Group 2

Group 2 to Group 1
Group 1 to Group 2
Group 2 to Group 3
Group 3 to Group 4.
Group 4 to Group 5.
Stop even if there are less than two respondents or if less than half of the mailed-and-delivered-to operators respond (Note: the last two steps are done only when less than half of the mailed-and-delivered-to operators in the left-hand group respond).

Screener respondents with screener-reported sales above the "must" cutoff become *musts*. A non-must screener respondent (from now on, a "screener respondent" means an in-scope screener respondent, where the screener itself has been used to determine in-scope status) is assigned to an NR group based on its list-frame country and sales category as determined on the screener.

Integerization

We integerize nonresponse weights using systematic probability sampling. This consists of the following steps:

1. Rewrite each coverage-adjusted weight as $a_k = a_{[k]} + r_k$, where $a_{[k]}$ is the largest integer less than or equal to a_k. In other words, $a_{[k]}$ is the integer portion of a_k, while r_k is its remainder.
2. Sort the *n* Census records in a state by *reporting* county, then by whether or not the record had crops or livestock on Indian-reservation land, and then by the Census-reported total land on each record. Without loss of generality, let us say that the records are already in this ordering so we can continue to use the same subscripts.
3. Choose a random *start point*, s, between 0 and 0.999999 from the uniform distribution.

4. For each Census record k, calculate the *cumulative sum* $t_k = r_1 + r_2 + ... + r_k$. Let $t_{[n]}$ denote the integer portion of t_n. By definition, $t_0 = 0$.
5. Call $H = \{h, h + 1, h + 2, ..., h + t_{[n]}\}$ the "set of *hit points*."
6. When cumulative sum, t_k, is greater than a hit point, say $h + j$, but t_{k-1} is not, set $a_k^{[I]} = a_{[k]} + 1$. Otherwise, set $a_k^{[I]} = a_{[k]}$.

(Note: it is possible for k to be the first Census record with $t_k > h + j$, while t_{k-1} is the first Census record with $t_{k-1} > h + j - 1$. In that case, the integerized weights for both k and $k - 1$ would be rounded up.)

The methodology described above assures that the integerized nonresponse weight for record k must be either $a_{[k]}$ or $a_{[k]} + 1$. Moreover, the probability of rounding up is proportional to r_k. Sorting records by county assures that integerization does not change the weighted number of farms in a county by more than 1. Sorting by land within county likewise assures that integerization does not change the weighted number of farms in a land-size category in a county by more than 1. It should also limit the amount by which integerization can change the weighted total land in a county.

A Simple Example of Integerization
(with s = 0.4)

Census Respondent	Nonresponse Weight	Cumulative Sum	Integerized Weights (see below)
1	1.01	.01	1
2	1.10	.11	1
3	1.01	.12	1
4	1.01	.13	1
5	1.02	.15	1
6	1.20	.35	1
7	1.01	.36	1
8	1.05	.41	2*
9	1.01	.42	1
10	1.01	.43	1
11	1.01	.44	1
12	1.02	.46	1
13	1.01	.47	1
14	1.01	.48	1
15	1.30	.78	1
16	1.30	1.08	1
17	1.02	1.10	1
18	1.01	1.11	1
19	1.01	1.12	1
20	1.04	1.16	1
21	1.01	1.17	1
22	1.10	1.27	1
23	1.05	1.32	1
24	1.01	1.33	1
25	1.01	1.34	1
26	1.30	1.64	2*
27	1.01	1.65	1
28	1.05	1.70	1
29	1.20	1.90	1
30	1.30	2.20	1

Since s = 0.4, the hit points are 0.4, 1.4, and 2.4. In this example, all the remainders are less than 0.5, but the weights of units 8 and 26 are rounded up. Note that if s were less than 0.2, there would have been three units with weights that rounded up.

Coverage Adjustment

The term “coverage-adjusted weight” is a misnomer because it includes nonresponse and other adjustments in addition to that for undercoverage of the census mail list. Nevertheless, we use it here.

The coverage-adjusted weights of *must* records are set at 1. The remaining records are adjusted by a complicated mathematical formula that is explained more fully elsewhere (Fetter and Kott, 2003, http://www.fcsm.gov/03papers/fetter_kott.pdf).

Each coverage-adjusted weight, w_k, falls in the range $1 \le w_k \le 6$. Usually, w_k is no less than the analogous nonresponse weight, a_k, but when $a_k > 1$, that need not be true.

Re-integerization

To limit the impact of integerization on the difference between aggregates computed with integerized coverage-adjusted and integerized nonresponse weights, the following routine, called “re-integerization,” was developed. The routine can be used whenever NASS needs to rerun the program that produces coverage-adjusted weights.

As described below,re-integerization is very similar to integerization except that a step, 2a, is added:

1. Rewrite each coverage-adjusted weight as $w_k = w_{[k]} + r_k'$, where $w_{[k]}$ is the largest integer less than or equal to w_k.
2. Sort the *n* Census records in a state by *reporting* county and then by their Census-reported total land on each record. Again, without loss of generality, let us say that the records are already in this ordering so we can continue to use the same subscripts.

2a. Define

$r_k'' = r_k'$ when $w_k \geq a_{[k]} + 1$ or $w_k \leq a_{[k]}$

$r_k'' = 0$ when $a_{[k]} < w_k < a_{[k]} + 1$, $r_k' < r_k$, and $a_k^{(I)} = a_{[k]}$

$r_k'' = r_k'/r_k$ when $a_{[k]} < w_k < a_{[k]} + 1$, $r_k' < r_k$, and $a_k^{(I)} = a_{[k]} + 1$

$r_k'' = (r_k' - r_k)/(1 - r_k)$ when $a_{[k]} < w_k < a_{[k]} + 1$, $r_k' \geq r_k$, and $a_k^{(I)} = a_{[k]}$

$r_k'' = 1$ when $a_{[k]} < w_k < a_{[k]} + 1$, $r_k' \geq r_k$, and $a_k^{(I)} = a_{[k]} + 1$.

3. Choose a random *start point*, s, between 0 and 0.999999 from the uniform distribution.

4. For each Census record k, calculate the *cumulative sum* $t_k = r_1'' + r_2'' + ... + r_k''$. Let $t_{[n]}$ denote the integer portion of t_n. By definition, $t_0 = 0$.

5. Call H = {h, h + 1, h + 2, ..., h + $t_{[n]}$} the "set of *hit points*."

6. When cumulative sum, t_k, is greater than a hit point, say h + j, but t_{k-1} is not, set $w_k^{[I]} = w_{[k]} + 1$. Otherwise, set $w_k^{[I]} = w_{[k]}$.

Drawing a systematic probability proportional to r_k'' sample and then rounding up or down as above creates a set of re-integerized nonresponse weights. It is unbiased because the probability of rounding w_k up when $w_k < a_{[k]} + 1$ is the sum of the probability of rounding a_k up (i.e., r_k) and the probability of *not* rounding a_k up but then selecting w_k to be rounded up (i.e., $[1 - r_k] \times [r_k' - r_k]/[1 - r_k]$).

There is a problem with this approach, however. Let S be the set of all *n* census-respondent operations in a county. The re-integerized coverage-adjusted total number of farms, $\sum_S w_k^{[I]}$, can differ from the total before integerization, $\sum_S w_k$ by more than 1. To ameliorate this situation, let S* be the subset of operations in a county containing those k for which $0 < r_k'' < 1$. Set

$r_k^* = r_k'' - [(\sum_S r_i'' - \sum_S r_i') / \sum_{S^*} r_i'']r_k''$ when $k \in S^*$ and $\sum_S r_i' \leq \sum_S r_i''$

$r_k^* = r_k'' + [(\sum_S r_i' - \sum_S r_i'') / \sum_{S^*} (1 - r_i'')](1 - r_k'')$ when $k \in S^*$ and $\sum_S r_i' > \sum_S r_i''$

$r_k^* = r_k''$ when $k \notin S^*$

Replacing the r_j'' in Step 4 by r_j^* is close to unbiased (asymptotically equivalent, in fact) while limiting the potential difference between $\sum_S w_k^{[I]}$ and $\sum_S w_k$.

Sample Weighting

A sample of Census records contain additional data items for select items either provided by the operator or imputed by NASS. See Appendix C, http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf, pp. C-1 & C-2.

Each sample record k has an expansion weight E_k. Sample weighting involves five steps:

A. *Nonresponse weighting* – a nonresponse weight for record k is denoted by a_k.

B. *Coverage adjustment* – a coverage-adjusted weight is denoted by w_k.

C. *Sample reweighting* – a sample weight is denoted by L_k.

D. *Integerization* – an integerized (coverage-adjusted) sample weight is denoted by $L_k^{(I)}$.

E. *Re-integerization* – an integerized nonresponse coverage-adjusted sample weight denoted (computed for analytical purposes).

Steps A and B have already been described. For Step C, let sample record, k have a sampling expansion of E_k. This value is 1 for certainties (records sampled with certainty are either musts or in certain list-frame counties) and some other value (2, 4, 6, or 8) for probabilities. A record's E_k value depends on its size on the frame and its *list-frame* county. Exactly how E_k is determined is explained elsewhere (see *Appendix C*, p. C-2).

In determining *sampling reweighting* (SR) groups for the post-stratification described below, we are concerned with a record's Census-reported TVP and the record's Census-reported county. Keep in mind, however, that E_k is determined by the list-frame county of k and its size on the frame.

We divide the sample in a reported county into SR groups based on . The largest-size group have size boundaries (based on Census-reported sales and land cutoffs) that vary by state. These cutoffs are the same as those employed for determining musts; now, however, Census values not frame values are used. Census records with TVP or land values in the range of the largest size group have sample data imputed for them if they were not selected for the sample. They are treated as sample records in what follows.

The SR groups as defined *within a reported county* as follows:

Group 0: The largest size group (includes all "musts" and all records in Alaska and Rhode Island)
Group 1: Other sample records (records not in Group 0) with TVS $\geq$ 150,000
Group 2: Other sample records with 150,000 > TVS $\geq$ 50,000
Group 3: Other sample records with 50,000 > TVS $\geq$ 10,000
Group 4: Other sample records with 10,000 > TVS $\geq$ 2,500
Group 5: Other sample records with TVS < 2,500.

Let CSR(k) denote the set of Census in-scope records in the same SR group as k and SSR(k) denote the set of sample in-scope records in the same SR group as k. We want to create sample weights that produce nearly unbiased estimates of the Census totals. This can be done by setting

$$L_k = w_k \left\{1 + (E_k - 1) \frac{\sum_{j \in CSR(k)} w_j - \sum_{j \in SSR(k)} w_j}{\sum_{j \in SSR(k)} w_j(E_j - 1)}\right\} \qquad \text{(A2.1)}$$

when at least one $E_j > 1$ for $j \in SSR(k)$

$$= w_k E_k \frac{\sum_{j \in CSR(k)} w_j}{\sum_{j \in SSR(k)} w_j E_j} \qquad \text{(A2.2)}$$

otherwise.

Note that the ratio in equation (A2.1) is always non-negative. Thus, $L_k \geq w_k \geq 1$.

Moreover, this ratio in (A2.1) or (A2.2) has as an expectation approximately equal to 1 when SSR(k) is large enough. Consequently, $L_k \approx w_k E_k$. The same is true trivially when all E_j in CSR(k) equal 1.

Observe that when the E_j in a group are identical, equation (A2.1) collapses to the more standard form:

$$L_k = w_k \{1 + \frac{\sum_{j \in CSR(k)} w_j - \sum_{j \in SSR(k)} w_j}{\sum_{j \in SSR(k)} w_j} \}.$$

$$= w_k \frac{\sum_{j \in CSR(k)} w_j}{\sum_{j \in SSR(k)} w_j} \ .$$

The same holds true for equation (A2.2).

We would like all the E_j in a group to be identical. This desire is not always realized because some groups contain records from more than one *list-frame* county, and list-frame counties can have different sampling rates.

We never collapse another group into Group 0 or *vice versa*. Other than that restriction, we collapse groups having small sizes in order for the estimators using equations (A2.1) or (A2.2) to be nearly unbiased. In what follows, an SR group is defined to be *small* if it contains less than six sample records OR if it has at least one but less than six sample records with $E_j > 1$ (we need at least six non-zero values within the summation in the denominator of (A2.1) and (A2.2)).

These steps are to be done in order:

1. If Group 1 is small, collapse its records into Group 2 (1 ceases to exist).
2. If Group 2 is small, collapse its records into Group 3 (2 ceases to exist).
3. If Group 3 is small, collapse its records into Group 4 (3 ceases to exist).
4. If Group 4 is small, collapse its records into Group 5 (4 ceases to exist).

5. If Group 5 is small, collapse its records into the next smallest group still existing *but not into Group 0*.
6. If Group x is small and the only remaining group other than Group 0, use it.

In invoking the last Step, we abandon near unbiasedness, because the sample in the county is too small for that property to have meaning. We still force the sample and census number of farms in the county to be equal. (This strict equality is lost, however, when the census and sample weights are integerized.)

We do not collapse into Group 0 because we do not want a record in that group ever to have a sample weight greater than 1, which could happen in a collapsed group when every e_j = 1 and equation (A2.2) is used.

Integerizing and re-integerizing Sample Weights

A nonresponse-adjusted-only sample weight, D_k, is calculated in an analogous manner to L_k in equations (A2.1) and (A2.2) with a_k (and a_j) replacing w_k (and w_j). These weights are integerized using the same methodology as the nonresponse weights – with D_k replacing a_k – and then re-integerized with L_k replacing w_k (and again D_k replacing a_k).

Smoothing State Farm Counts for the 2002 Census of Agriculture

The direct estimates of the 2002 farms counts (census + nonresponse + NML) for many states were statistically unreliable. This note first describes one possible method of smoothing those estimates to be closer to previously published NASS values, called "Board numbers." An alternative method, described afterwards, was used for the 2002 Census of Agriculture with only small modification.

Let i denote a state (or New England treated like a single entity), and let B_i be the Board value for the number of farms in the state. Let T_i be the farm count for the state based on the Census of Agriculture (respondent-adjusted list plus NML), hereafter called the "Census number."

A potential smoothed estimate for the farm count in State i is

$$M_i = (1 - \lambda_i)T_i + \lambda_i B_i (\sum \lambda_j T_j / \sum \lambda_j B_j), \qquad \text{(B1)}$$

where the summations are over all states. Determining the smoothing factor, λ_j, for each state j, a value between 0 and 1, is the centerpiece of this note.

Equation (B1) has several nice properties:

A. $\sum M_i = \sum T_i$, no matter the choice for the λ_i.

B. At one extreme (when $\lambda_i = 0$), $M_i = T_i$.

C. At the other extreme (when $\lambda_i = 1$), $M_i = B_i (\sum \lambda_j T_j / \sum \lambda_j B_j)$, which is close to B_i when $\sum \lambda_j T_j / \sum \lambda_j B_j$ is close to 1.

The smoothing factors can be determined ideally under a model like

$$E(T_i)/B_i = \mu + \delta_i, \qquad \text{(B2)}$$

where the δ_i are independent random variables with mean zero and variance σ^2. The value $E(T_i)$ is unknown. It is the expectation of T_i with respect the sampling done when estimating its NML component (the expectation can "average out" nonsystematic measurement-error within each state).

In this first setup, B_i is treated as fixed, and the goal is to estimate $\theta_i = E(T_i) = B_i(\mu + \delta_i)$ for each i. Note that δ_i is formally a random component, not an error. Effectively, however, it is a quantification of the random, state-level measurement error in B_i, while the difference between μ and 1 quantifies the systematic measurement error across all the B_i.

Suppose we knew σ^2 and the sampling variance for each T_i (which we assume comes entirely from the NML). Call the latter V_i^2. Then the ideal value for λ_i is

$$\lambda_i^{IDEAL} \approx V_i^2/(V_i^2 + B_i^2\sigma^2).$$

In practice, V_i^2 can be replaced by an estimate of the variance of T_i. Call it v_i^2. It is helpful to denote the pseudo-CV of t_i, v_i/B_i as cv_i ("pseudo" because the division is by B_i instead of T_i). Given an estimator for σ^2, call it s^2, a good choice for λ_i is

$$g_i = cv_i^2/(cv_i^2 + s^2). \qquad \text{(B3)}$$

An unbiased estimator for σ^2 is

$$s^2 = \{ \sum (T_i/B_i)^2 - [\sum (T_i/B_i)]^2/n\}/(n - 1) - \sum cv_i^2/n. \qquad \text{(B4)}$$

The Preferred Alternative

In the preferred alternative method, a state (or New England) Board number, B_i, is still assumed to be a fixed predictor of the true number of farms in State i, $\theta_i = B_i(\mu + \delta_i)$ (see the discussion surrounding equation (B2)), where δ_i is effectively a quantification of the random state-level measurement error of the Board number.

The sampling expectation of a Census number, T_i, averages out the effects of unit-level nonsystematic measurement error with state i. There remains, however, a potential for systematic measurement error in $E(T_i)$.

The alternative method described below assumes that state-level measurement error in $E(T_i)$ has zero mean; that is, it “averages out” across the states. Moreover, it has the same variance as δ_i (i.e., it is no more precise than B_i). Formally, the assumed model is

$$E(T_i)/B_i = \mu + \delta_i + \epsilon_i,$$

where $E(\delta_i) = E(\epsilon_i) = 0$, $E(\delta_i^{\,2}) = E(\epsilon_i^{\,2}) = \sigma^2/2$, and the target of estimation is $B_i(\mu + \delta_i)$.

Equation (B4) again provides an unbiased estimator for σ^2, while a good choice for λ_i is now

$$g_i = (cv_i^{\,2} + s^2/2)/(cv_i^{\,2} + s^2). \qquad \text{(B5)}$$

Although it is possible to estimate the variance of $M_i = (1 - g_i)T_i + g_i B_i\,(\sum g_jT_j / \sum g_jB_j)$ under this model and choice for g_i, a prudent course of action, which was largely adopted, is to use the estimated variance for T_i as a conservative indication of the variance for M_i.

Since NASS is confident that none of its estimated state farm counts are off by more than 10%, the largest the estimated variance for M_i was allowed to be was $(.01)M_i^2$. That is to say, NASS uses $\max\{v_i^2, (.01)M_i^2\}$ as the variance estimator for M_i in the 2002 Census of Agriculture.

A Components-of-Variance-Model Estimator for State-Level NML Proportions

This note describes the estimator for a state-level NML proportion like the fraction of farm operations with horses. By multiplying this proportion by the (smoothed) estimate of NML farms in a state and adding it to the nonresponse-adjusted census total, NASS computed the calibration target for the number of farms in the state with horses.

Let i denote a state (i = 1, .., T) and k an NML farm within the state (New England is treated as a state). If y_{ik} is the 0/1 item value of interest for farm k in state i, and W_{ik} is the farm's sampling weight (including the tract-to-farm ratio), then the usual estimator of the state NML proportion for the item (or, more precisely, of the fraction of farms having an item value of 1) is

$$y_i = \sum_{k \in S(i)} W_{ik}\, y_{ik} \,/\, \sum_{k \in S(i)} W_{ik} \;=\; \sum_{k \in S(i)} w_{ik}\, y_{ik}, \tag{C1}$$

where S(i) is the NML sample of farms in state i, and $w_{ik} = W_{ik} / \sum_{h \in S(i)} W_{ih}$,

For most of this note, we ignore the fact that some y_{ik} are imputed with values between 0 and 1. When we finally estimate the variance of the proposed composite estimator of a proportion, we adjust for this ignorance.

The usual estimator for the US-level NML proportion is

$$y = \sum^T W_i\, y_i \,/\, \sum^T W_i\,, \tag{C2}$$

where $W_i = \sum_{k \in S(i)} W_{ik}$. We assume here that the mean sqared error of y, unlike y_i, is acceptably low.

Let c_i be the estimated proportion of the item in the state derived from the Census list (adjusted for nonresponse). It is convenient to define $d_i = y_i / c_i$ (which means we have to remove any state with $c_i = 0$ from T) and rewrite y_i as $y_i = c_i\, d_i$.

We focus our attention on an estimator for the state NML proportion of the form

$$z_i = c_i\,[(1 - \lambda_i)d_i + \lambda_i \textstyle\sum^T f_j\, d_j]$$

$$= c_i\,[(1 - \lambda_i)d_i + \lambda_i\, d]\ , \tag{C3}$$

where $d = \sum^T f_j\, d_j$, and the f_j are arbitrary factors that sum to 1 (i.e., $\sum^T f_j = 1$).

When $\lambda_i = 1$, equation (C3) replaces the state-specific estimator of the ratio between the NML and census proportions, d_i, with a pooled estimator of this ratio, d. When λ_i takes on a value between 0 and 1, the equation compromises between using d_i and d.

Later, we choose specific values for the f_j that have useful properties. We cannot compute d until we choose values for the f_j. With this in mind let

$$d^{(0)} = \textstyle\sum^T w_i d_i\ ,$$

where $w_i = W_i c_i / \sum^T W_j c_j$.

The Components-of-Variance Model

In order to determine a good value for λ_i, we posit this components-of-variance (or random-effects) model for the $d_{ik} = y_{ik}/c_i$:

$$d_{ik} = \mu + \eta_i + \epsilon_{ik}\ ,$$

where $E(\eta_i) = E(\epsilon_{ik}) = 0$, and all the η_i and ϵ_{ik} are uncorrelated, $Var(\epsilon_{ik}) = \sigma_i^2$, and $Var(\eta_i) = \sigma_B^2$. The interested reader will observe that this model treats the d_{ik} within state i as random variables with a common mean, $\mu + \eta_i$. These state means themselves have a common mean, μ, and variance, σ_B^2. Note that it is the ratio between the state NML and census proportions of the item of interest that is being modeled rather than the NML proportions themselves. Modeling ratios in this way is statistically identical to modeling percentage differences.

Our goal is to estimate $Y_i = \sum_{U(i)} y_{ik}/N_i$, where U(i) is the set of all N_i NML farms, whether or not in the sample, in state i. Observe that $Y_i = c_i D_i$, where

$D_i = \sum_{U(i)} d_{ik}/N_i$. For all practical purposes, $D_i = \mu + \eta_i + \sum_{U(i)} \epsilon_{ik}/N_i$ can be approximated by $D_i \approx \mu + \eta_i$. Thus, the (model) variance of $d_i = \sum_{S(i)} W_{ik} c_i d_{ik} / \sum_{S(i)} W_{ik} c_i = \sum_{S(i)} w_{ik} d_{ik}$ as an estimator for D_i is

$$E[(d_i - D_i)^2] \approx (\textstyle\sum_{S(i)} w_{ik}^2) \sigma_i^2$$

$$= V_i \text{ (this defines } V_i).$$

Generally, V_i decreases as the sample size of S(i) increases.

We treat the y_{ik} as if they were generated by a Bernoulli process with mean Y_i. Consequently, $\mathrm{Var}(y_{ik} \mid Y_i) = Y_i(1 - Y_i)$, $\mathrm{Var}(\epsilon_{ik}) = Y_i(1 - Y_i)/c_i^2$, and

$$V_i = (\textstyle\sum_{S(i)} w_{ik}^2) Y_i(1 - Y_i)/c_i^2.$$

Rather than dealing with the model variance of d as an estimator for D, it is more convenient to examine the properties of

$$d_{(i)} = \textstyle\sum_{j \neq i} f_j d_j /(1 - f_i),$$

which has mean μ and is uncorrelated with both d_i and D_i. This random variable relates to d and d_i through

$$d = f_i d_i + (1 - f_i) d_{(i)} .$$

It is also useful to observe that $d_i - d = (1 - f_i)(d_i - d_{(i)})$.

The estimator z_i in (C3) can be rewritten as

$$z_i = c_i [(1 - \lambda_i^*) d_i + \lambda_i^* d_{(i)}], \qquad \text{(C3*)}$$

where $\lambda_i^* = \lambda_i(1 - f_i)$. We use this later.

The variance of $d_{(i)}$ as an estimator of D_i is

$$E[(d_{(i)} - D_i)^2] = \textstyle\sum_{j \neq i} f_j^2(V_j + \sigma_B^2)/(1 - f_i)^2 + \sigma_B^2.$$

In practice, we do not know the values for the V_i^2 and σ_B^2 and need to estimate them from the sample. Assuming (initially) that each Y_i is approximately equal to $c_i d^{(0)}$, this can be done with

$$v_i^{(0)} = \sum_{S(i)} w_{ik}^2\, c_i d^{(0)} (1 - c_i d^{(0)})/c_i^2, \text{ and}$$
$$b \;=\; [\sum^T w_i\, (d_i - d^{(0)})^2 - \sum^T w_i\, (1 - w_i) v_i^{(0)}] \,/\, [1 - \sum^T w_i^2]. \qquad \text{(C4)}$$

Choosing Parameters

An obvious thing to do is to choose the f_j and λ_j so that the variance for each z_i is minimized. Given a set of f_j, we have

$$\mathrm{Var}(z_i) = c_i^{\,2}\{ (1 - \lambda_i^*)^2 V_i + (\lambda_i^*)^2 \sum_{j \neq i} f_j^2 (V_j + \sigma_B^2)/(1 - f_i)^2 + \sigma_B^2. \qquad \text{(C5)}$$

Setting the derivative of the right-hand size of equation (C5) with respect to λ_i^* equal to 0 and then solving for the optimal λ_i^* yields:

$$\lambda_i^*{}_{OPT} = V_i \,/\, [V_i + \sum_{j \neq i} f_j^2 (V_j + \sigma_B^2)/(1 - f_i)^2 + \sigma_B^2].$$

If all the f_j are small, and σ_B^2 is not too small, then

$$\lambda_{i\ A.OPT} \approx \lambda_i^*{}_{A.OPT} \approx V_i \,/\, [V_i + \sigma_B^2] \qquad \text{(C6)}$$

is approximately optimal no matter what the choice for the f_j.

One property we would like the z_i to have is that their weighted mean, $\sum^T W_i z_i / \sum^T W_i$, equal $y = \sum^T W_i y_i / \sum^T W_i$, because, at the aggregate level, y is a good estimate. Given any set of λ_j, this requirement when applied to equation (C3) forces

$$f_i = W_i c_i \lambda_i \,/\, \sum^T W_j c_j \lambda_j. \qquad \text{(C7)}$$

In practice, we can implement neither equation (C6) nor (C7) because the V_i and $\sigma_B{}^2$ are unknown. Instead, they are estimated with

$v_i{}^* = v_i{}^{(0)}$,

$b^* = \max\{b,\ \tfrac{1}{2} \sum^T w_i\,(d_i - d^{(0)})^2\}$,

so that equations (C6) and (C7) become

$$\ell_i \approx v_i{}^* / [v_i{}^* + b^*], \quad \text{and} \tag{C6'}$$

$$f_i = W_i c_i\, \ell_i / \sum^T W_j c_j\, \ell_j. \tag{C7'}$$

Consequently, this chosen version of z_i

$$z_i{}^C = (1 - \ell_i) y_i + \ell_i (c_i d) \tag{C8}$$

is approximately optimal.

Variance Estimation and Confidence Intervals

We now discuss variance estimation for $z_i{}^C$ and confidence interval construction for Y_i. First, we put $z_i{}^C$ in a form that simplifies the variance derivation:

$$z_i{}^C = (1 - \ell_i[1 - f_i]) y_i + \ell_i[1 - f_i] c_i d_{(i)}.$$

Treating the ℓ_i as fixed, the variance of $z_i{}^C$ is roughly:

$$\mathrm{Var}(z_i{}^C) = (1 - \ell_i[1 - f_i])^2\, V_i + (\,\ell_i[1 - f_i] c_i)^2\ \{\, \textstyle\sum_{j\neq i} f_j^2 (V_j + \sigma_B{}^2)/(1 - f_i)^2 + \sigma_B{}^2 \}.$$

We can again estimate σ_B^2 with b or b* as appropriate. For V_j, however, we now suspect that z_j^C is a much better guess at Y_j than $c_i d$. Consequently, we can compute

$$v_j^{(1)} = (\sum_{S(j)} w_{jk}^2) z_j^C (1 - z_j^C)/c_j^2.$$

At this point, let us offer an *ad hoc* adjustment for the fact that some of the y_{ik} may be imputed. Let $\omega_{jk} = w_{jk}R_{jk}/ \sum_{h \in S(j)} w_{jh}R_{jh}$, where $R_{jh} = 1$ when y_{jh} is a reported value and $R_{jh} = 0$ when it is imputed. We replace $v_j^{(1)}$ above with

$$v_j^{(1)} = (\sum_{S(j)} \omega_{jk}^2) z_j^C (1 - z_j^C)/c_j^2. \qquad \text{(C9)}$$

Equation (C9) leads to the variance estimator:

$$\text{var}(z_i^C) = (1 - \ell_i[1 - f_i])^2 c_i^2 v_i^{(1)} + (\ell_i[1 - f_i]c_i)^2 \{ \sum_{j \neq i} f_j^2(v_j^{(1)} + b)/(1 - f_i)^2 + b^* \}. \qquad \text{(C10)}$$

We can construct a two-sided, Wilson-like confidence interval for Y_i by solving

$$\frac{(z_i^C - Y_i)^2}{(1 - \ell_i[1 - f_i])^2 c_i^2 V_i + (\ell_i[1 - f_i]c_i)^2 \{ \sum_{j \neq i} f_j^2(v_j^{(1)} + b)/(1 - f_i)^2 + b^* \}} \leq t^2, \qquad \text{(C11)}$$

for Y_i, where t is the relevant z-score (i.e., 1.96 for a two-sided 95% confidence interval). The left hand side of equation (C11) is the *square* of the Wald pivotal for z_i^C as an estimator for Y_i except that V_i replaces $v_i^{(1)}$ (note: when x is an unbiased estimator for X, the Wald pivotal for x is $[x - X]/\sqrt{\text{var}[x]}$).

Recall that $c_i^2 V_i = (\sum_{S(i)} w_{ik}^2) Y_i(1 - Y_i)$. For convenience, we set

$\alpha_i = (1 - \ell_i[1 - f_i])^2 \sum_{S(i)} w_{ik}^2$, and

$\beta_i = (\ell_i[1 - f_i]c_i)^2 \{\sum_{j \neq i} f_j^2 (v_j^{(1)} + b)/(1 - f_i)^2 + b^*\}$, so that $\mathrm{var}(z_i^C) = \alpha_i\, z_i^C\,(1 - z_i^C) + \beta_i$. Equation (C11) can now be rewritten as

$$\frac{(z_i^C - Y_i)^2}{\alpha_i Y_i(1 - Y_i) + \beta} \le t^2 \quad \text{or}$$

$$(z_i^C)^2 - 2\, z_i^C Y_i + Y_i^2 - t^2\, \alpha_i Y_i + t^2\, \alpha_i Y_i^2 - t^2\, \beta_i \le 0$$

$$\{1 + t^2\alpha_i\}Y_i^2 - \{2\, z_i^C + t^2\, \alpha_i\}Y_i + \{(z_i^C)^2 - t^2\, \beta_i\} \le 0. \qquad \text{(C12)}$$

Solving equation (C12) when equality holds yields (recall that if $Ax^2 + Bx + C = 0$, then $x = [-B \pm \surd\{B^2 - 4AC\}] / [2A]$):

$$Y_i = \frac{2\, z_i^C + t^2\alpha_i \pm [\{2\, z_i^C + t^2\alpha_i\}^2 - 4\,\{1 + t^2\alpha_i\}\{(z_i^C)^2 - t^2\, \beta_i\}]^{1/2}}{2\,\{1 + t^2\alpha_i\}}$$

$$= \frac{z_i^C + t^2\alpha_i(½) \pm [\{\, z_i^C + t^2\alpha_i/2)\}^2 - \{1 + t^2\alpha_i\}\{(z_i^C)^2 - t^2\, \beta_i\}]^{1/2}}{1 + t^2\alpha_i}$$

$$= \frac{z_i^C + t^2\alpha_i(½) \pm [z_i^C\,(1 - z_i^C)]t^2\, \alpha_i + t^4\, \alpha_i^{\,2}/4 + \{1 + t^2\alpha_i\}t^2\, \beta_i\}]^{1/2}}{1 + t^2\alpha_i}$$

$$= \frac{z_i^C + t^2\alpha_i(½)}{1 + t^2\alpha_i} \pm \frac{t\,[\mathrm{var}(z_i^C) + t^2\alpha_i^{\,2}/4 + t^2\alpha_i\beta_i]^{1/2}}{1 + t^2\alpha_i} \qquad \text{(C13)}$$

Equation (C13) defines the endpoints of a two-sided confidence interval for Y_i. It is interesting to note that the center of this interval (the first term on the right) is a weighted average of z_i^C and ½. That is to say, when $t^2\alpha_i < 1$, which is almost certainly the case, the center will between z_i^C, the standard center of a two-sided interval and ½.

The second term on the right in equation (C13) is likely to be close to var(z_i^C).

Categories

So far in this paper we have estimated related fractions, like those of the NML operations in particular sales classes, independently of each other. As a result, if there are G mutually exclusive and exhaustive categories (like sales classes or age groups) in a state, then the estimated fractions across the category need not sum to 1. That is to say, when z_{gi}^C is the estimate for the fraction of NML farms in category g (=1, ..., G) derived from equation (C8) (with the subscript g added to denote the category), there is no reason for $\sum^G z_{gi}^C$ to equal 1.

The following iterative method should produce revised fractions, z_{gi}^R, that sum both ways; that is, across category (g): $\sum^G z_{gi}^R = 1$, and across states (i): $\sum^T W_i z_{gi}^R = \sum^T W_i y_{gi}$. We begin by computing $\ell_{gi}^{(1)}$ and $f_{gi}^{(1)}$ using equations (C6') and (C7'), respectively for each category g (again adding the subscript g as appropriate). Let

$$d_{g+}^{(r)} = \sum_{i=1}^{T} f_{gi}^{(r)} d_{gi}. \qquad \text{(C14)}$$

for a particular set, where r can be 1, 2, etc. In addition, let

$$\begin{array}{llll} \ell_{gi}^{(r+1)} & = & L_{gi}^{(r+1)} & \text{if } \ 0 \le L_{gi}^{(r+1)} \le 1 \\ & = & 1 & \text{if } \ L_{gi}^{(r+1)} > 1 \\ & = & 0 & \text{if } \ L_{gi}^{(r+1)} < 0, \end{array} \qquad \text{(C15)}$$

where

$$L_{gi}^{(r+1)} = \ell_{gi}^{(r)} - \{[\sum_{h=1}^{G} \ell_{hi}^{(r)}(y_{hi} - c_{hi}d_{h+}^{(r)})] / \sum_{h=1}^{G} |y_{hi} - c_{hi}d_{h+}^{(r)}|\} \operatorname{sgn}(y_{gi} - c_{gi}d_{g+}^{(r)}),$$

and sgn(x) = 1 when x > 0, –1 when x < 0, 0 otherwise.

And

$$f_{gi}^{(r+1)} = W_i c_{gi}\, \ell_{gi}^{(r+1)} / \sum_{j=1}^{T} W_j c_{gj}\, \ell_{gj}^{(r+1)}. \tag{C16}$$

After several iterations (say r = 4), we have

$$z_{gi}^{R} = (1 - \ell_{gi}^{(r)}) y_{gi} + \ell_{gi}^{(r)}(c_{gi} d_{g+}^{(r)}) = y_{gi} + \ell_{gi}^{(r)}(c_{gi} d_{g+}^{(r)} - y_{gi}).$$

Equation (C14) calculates the d_{g+} is such a way that the z_{gi}^{R} sum across states (i.e., $\sum^{T} W_i z_{gi}^{R} = \sum^{T} W_i y_{gi}$ for each g). Equation (C15) forces the z_{gi}^{R} to sum to 1 across categories (for each state i). Iteration may be necessary because doing one can undo the other. Equations (C14) and (C16) are mild generalization of previous formulae. Equation (C15) attempts to minimize the largest absolute change among the ℓ_{gi} across iterations in such a way that the z_{gi}^{R} sum to 1 across classes, while all the l_{gi} remain between 0 and 1.

We can estimate the variance of z_{gi}^{R} and compute its confidence interval using the same methods we developed for z_i^{C} substituting z_{gi}^{R} and $l_{gi}^{(r)}$ (for sufficiently large r) as appropriate in equations (C9) through (C13).

Covariance Estimation

Let P be the number of fractions for which NASS estimates proportions to use in its calibration program. Let z_{pi}^{C} (or z_{pi}^{R}) represent the estimator for the pth fraction in state i, and y_{pik} be the p value (reported or imputed) for farm k in state i.

We define

$$\rho_{pqi} = \sum_{k \in S(i)} w_{ik}(y_{pki} - y_{pi})(y_{qki} - y_{qi}) / \{\sum_{k \in S(i)} w_{ik}(y_{pki} - y_{pi})^2 \sum_{k \in S(i)} w_{ik}(y_{qki} - y_{qi})^2\}^{½},$$

where $y_{pi} = \sum_{k \in S(i)} w_{ik} y_{pki}$.

A reasonable, if *ad hoc*, estimator for the covariance between z_{pi}^{C} and z_{qi}^{C} (replacing C by R when the fraction uses the revised estimate) is

$$\mathrm{cov}(z_{pi}^{C}, z_{qi}^{C}) = \rho_{pqi} \{\mathrm{var}(z_{pi}^{C})\mathrm{var}(z_{pi}^{C})\}^{½}, \tag{C17}$$

where $\mathrm{var}(z_{pi}^{C})$ is defined by equation (C10). Note that when p = q, $\mathrm{cov}(z_{pi}^{C}, z_{qi}^{C}) = \mathrm{var}(z_{pi}^{C})$.

Mean-Squared-Error Estimates for the 2002 Census of Agriculture

This note describes how mean squared errors are estimated in all states except Alaska and Hawaii. Weights in neither of these states compensate for coverage errors. Only Hawaii has weights adjusted for nonresponse. A discussion of how mean squared errors are calculated in Hawaii is reserved for the end of this note.

There are many sources of error in the 2002 Census of Agriculture. We focus here on creating combined mean-squared-error measures for (up to) four of those:

A. Reweighting for nonresponse,

B. Coverage adjustment,

C. Integerization, and

D. Sampling.

The last source of error applies only to sample numbers. In restricting ourselves to these four, we assume that all the data on all Census records are correct, including those from secondary sources and imputation. The measures discussed here are, mean-squared-error (mse) estimates rather than variances because they capture the additional error arising from integerization and (when deemed appropriate) using biased calibration targets.

Effectively, NASS adjusts for measurement error in the calibration process while it adjusts for undercoverage. Nevertheless, we use the term “coverage-adjusted” number here. In the formulae discussed here, we ignore both the contribution to mse caused by measurement error and the potential reduction of mse due to calibration on quantitative commodity and land-in-farms targets.

Three Types of Census Numbers

There are three different types of Census numbers of concern to us here. Each requires a different approach. The types are

1. Nonresponse-adjusted Census numbers (pre-integerized)
2. Coverage-adjusted Census numbers
3. Coverage-adjusted sample numbers

Nonresponse-Adjusted Census Numbers (before Integerization)

Let U_g (or Ug) denote a nonresponse group (g = 1, ..., G), and let r_g denote the number of Census respondents in g (r_g counts both in-scope and out-of-scope records). The variance/mean-squared error (mse) estimate for $t_y = \sum_U a_k y_k$ has the form:

$$\mathrm{mse}_N(t_y) = \sum^G (1 - [1/a_{(g)}])[r_g/(r_g - 1)][\, \sum_{k\in Ug} (a_k y_k)^2 - (\sum_{k\in Ug} a_k y_k)^2/r_g], \qquad \text{(D1)}$$

where

U is the set of *n* Census records in the state,

y_k is the value of interest for record k,

a_k is the nonresponse weight for record k before integerization, and

$a_{(g)}$ is the common value of a_k for all records in nonresponse group g.

The expression on the right of equation (D1) is the standard variance estimator under the all-form-recipients-within-a-group-are-equally-likely-to-respond assumption, although not in the form one usually sees it.

For a county-level aggregate (e.g., the total land in county q), y_k is defined to be positive only when k is in the county of interest. When estimating a farm count, y_k is defined to be either 0 or 1. For convenience, we define x_k =1 for all in-scope records, so that $t_x = \sum_U a_k x_k$ is the nonresponse-adjusted estimate of the number of farms in the state (before integerization).

Coverage-adjusted Census Numbers

Subsequent to the initial run of the calibration program that created the coverage-adjusted Census weights in each state, there were a number of data fixes. After that, a second run of the program was undertaken treating the calibration weights from the first run as the nonresponse weights. The quasi-randomization theory underlying the mean-squared error estimation described below pretends there was only one run of the calibration programs.

Suppose we have a coverage-adjusted Census total for a calibration state of the form $t_y^C = \sum_U w_k^{(I)} y_k$, where $w_k^{(I)}$ is the integerized coverage-adjusted weight for record k. Likewise, w_k is the coverage-adjusted weight before integerization.

Let P be the number of demographic variables *actively* targeted in *either* run of the calibration program when computing the coverage adjusted weights. The calculation of P *excludes* the simple count variable, which is 1 for all farms. It also excludes those demographic variables which were deemed in range without specific targeting in both runs of the calibration program.

For each farm k, let $\mathbf{z}_k$ denote a row vector of Q demographic calibration variables associated with k *not counting* the indicator variable for “extreme operators” or EO’s. These Q variables are often be the P actively targeted demographic variables; however, we need to drop one economic-size variable if all such variables are targeted and one age variable if all such variables are targeted (to avoid a singularity).

Let $\mathbf{g}_k = (1, \mathbf{z}_k)$, and $\mathbf{Z}_T$ be the targeted sum of the $\mathbf{z}_k$; that is, the vector of *original* targeted totals for the demographic variables. Furthermore, let $\mathbf{z}_T = \mathbf{Z}_T / X$ be the vector of proportions associated with $\mathbf{Z}_T$, where X is the targeted number of farms.

Sometimes, the calibration programs use a target that is at the boundary of the acceptable range. The components of $\mathbf{Z}_T$ are all original values and not boundary targets.

We now define

$$\mathbf{b} \;=\; \left(\textstyle\sum_{U^*} a_k \mathbf{g}_k{}' \, \mathbf{g}_k \right)^{-1} \textstyle\sum_{U^*} a_k \mathbf{g}_k{}' \, y_k = (b_0, \mathbf{b}_z{}')', \qquad \text{(D2.1)}$$

$$e_k \;=\; (y_k - \mathbf{g}_k \mathbf{b})(n_*/[n_* - Q - 1])^{½}, \qquad \text{(D2.2)}$$

where

U* the subset of U containing only records, k, such that $w_k > 1$ and $w_k < 6$, and

n_* the number of records in U*.

The $(n_*/[n_* - Q - 1])^{½}$ factor in the definition of the residual, e_k, is an *ad hoc* compensation for **b** being an estimated value.

We *could* estimate the contribution to the mse of t_y^C caused by coverage-adjustment (last line) and NML estimation of calibration targets (first two lines) as

$$\begin{aligned} mse_C(t_y^C) \;=\;& var_{NML\_x}(b_0 + \mathbf{z}_T \mathbf{b}_z)^2 + \\ & X_{NML}{}^2 \mathbf{b}_z{}' \, \mathbf{COV} \, \mathbf{b}_z + X^2(\mathbf{bias} \, \mathbf{b}_z)^2 \\ & \textstyle\sum_U w_k (w_k - a_k)_{w \geq a} \, e_k{}^2, \end{aligned} \qquad \text{(D3)}$$

where

X_{NML} is the estimated total of farms in the NML,

X is the total number of farms in the state,

var_{NML_x} is the estimated variance of the total farm number estimate from the NML,

bias = $(\sum_U w_k \mathbf{z}_k / \sum_U w_k) - \mathbf{z}_T$, and

COV is the Q x Q matrix whose p,qth term is the estimated covariance of smoothed proportion estimators described in equation (C17).

$(w_k - a_k)_{w>a} = w_k - a_k$ when $w_k \geq a_k$, 1 otherwise.

A farm may be on the Census list more than once without our knowing it. When $a_k/w_k \leq 1$, we estimate the variance of the number of times farm k is on the list with $(a_k/w_k)[1 - (a_k/w_k)]$. When $a_k/w_k > 1$, we estimate that variance conservatively with a_k/w_k. In the former case, we assume that the number of times farm k is on the list cannot exceed 1, and a_k/w_k is simply the probability k is on the list. In the latter case, we assume that the number of times k is on the list is a random variable with mean np and variance np(1 - p), where n and p are unknown, but np is estimated by a_k/w_k, so np(1 - p) which is bound above by np, is conservatively estimated by a_k/w_k.

Also on the conservative side, equation (D3) ignores quantitative calibration targets such as farm land and whatever mean-squared-error reducing power these targets provide. In addition, the smoothed number-of-farm targets used in calibration should have less variance than $\text{var}_{\text{NML\_x}}$.

There is a another term in the estimation of the mse of t_y^C contributed by the adjustment for nonresponse. It is

$$\text{mse}_N(t_y^C) = \sum^G (1 - [1/a_{(g)}])[r_g/(r_g - 1)][\sum_{k\epsilon Ug} (a_k y_k^C)^2 - (\sum_{k\epsilon Ug} a_k y_k^C)^2/r_g], \qquad \text{(D4)}$$

where $y_k^C = y_k + [(w_k/a_k) - 1]e_k$. Part of $\text{mse}_N(t_y^C)$ (the part associated with $y_k - e_k$) is the contribution of mse due to the list component of the estimated calibration target vector. The rest captures the impact of nonresponse weighting on the residual before coverage adjustment.

The total mse for an integerized coverage-adjusted Census number, $t_y^{C(I)} = \sum_U w_k^{(I)} y_k$ can be estimated with

$$\mathrm{mse}_T(\ t_y^{C(I)}) = \mathrm{mse}_C(\ t_y^{C(I)}) + \mathrm{mse}_N(t_y^{C}), \qquad \text{(D5)}$$

where $\mathrm{mse}_C(\ t_y^{C(I)})$ is modified from $\mathrm{mse}_C(\ t_y^{C})$ in equation (D3) to capture the impact of integerization like so:

$$\begin{aligned} \mathrm{mse}_C(t_y^{C(I)}) \quad = \quad & \mathrm{var}_{NML\_x}(b_0 + \mathbf{z}_T\mathbf{b}_z)^2 + \\ & X_{NML}{}^2\mathbf{b}_z' \ \mathbf{COV}\ \mathbf{b}_z + X^2(\mathbf{bias}_{(I)}\mathbf{b}_z)^2 \\ & \textstyle\sum_U w_k{}^{(I)}(w_k{}^{(I)} - a_k)_{w \geq a}\, e_k{}^2, \end{aligned} \qquad \text{(D3}^{(I)}\text{)}$$

where

$\mathbf{bias}_{(I)} \quad = \ (\ \sum_U w_k{}^{(I)}\ \mathbf{z}_k \,/\, \sum_U w_k{}^{(I)}) - \mathbf{z}_T$, and

$(w_k{}^{(I)} \ - a_k)_{w>a} = \ w_k{}^{(I)} \ - a_k$ when $w_k \geq a_k$, 1 otherwise,

The contribution of coverage to total mse adjustment, including for practical purposes, integerization is

$$\begin{aligned} R_C &= \mathrm{mse}_C(\ t_y^{C(I)})/\mathrm{mse}_T(\ t_y^{C(I)}) \\ &= \mathrm{mse}_C(\ t_y^{C(I)})/[\mathrm{mse}_C(\ t_y^{C(I)}) + \mathrm{mse}_N(t_y^{C})], \end{aligned} \qquad \text{(D6)}$$

If the right-hand side of equation (D6) is negative, R_C is changed to zero. See the appendix of this note.

The computed values for R_C as described above appear in percentage form in the last column of Table B in

http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf

and its state-level analogues,* which we have called *Appendix C*. The contribution of nonresponse adjustment to total estimated mse – also displayed in percentage form on Table B – is $(1 - R_C)$.

Coverage-adjusted Sample Numbers

Coverage-adjusted sample numbers like $t_y^S = \sum_S L_k^{(I)} y_k$, where S is the set of sample records in a state, and $L_k^{(I)}$ is the integerized sample weight for k, have three mse components: the original nonresponse adjustment, the coverage adjustment and the sample adjustment.

Since y_k is unknown for census records not in the sample, when estimating the contribution to mse of coverage adjustment and NML estimation, we modify equations (D2) and (D3) like so:

$$\mathbf{b}^S = \left(\sum_{S^*} a_k E_k \mathbf{g}_k' \, \mathbf{g}_k \right)^{-1} \sum_{S^*} a_k E_k \mathbf{g}_k' \, y_k = (b_x{}^S, \mathbf{b}_z{}^{S}{}')', \qquad \text{(D2.1}^S\text{)}$$

$$e_{kS} = (y_k - \mathbf{g}_k \mathbf{b}^S)(n_S/[n_S - Q - 1])^{1/2}, \qquad \text{(D2.2}^S\text{)}$$

For a state-level analogue replace the repeated 'us' in 'us/us2' with the appropriate state abbreviation; for example, the URL of Appendix C for Nebraska is http://www.nass.usda.gov/census/census02/volume1/ne/ne2appxc.pdf

$$
\begin{aligned}
\mathrm{mse}_C(t_y^S) \quad = \quad & \mathrm{var}_{NML\_x}(b_x + \mathbf{z}_T\mathbf{b}_z^S)^2 + \\
& X_{NML}^{2}\mathbf{b}_z^{S}{}'(\mathbf{COV} + \mathbf{BIAS}^2)\mathbf{b}_z^S + \\
& \textstyle\sum_S E_k w_k (w_k - a_k)_{w\geq a}\, e_{kS}^{2}, \qquad\qquad (D3^S)
\end{aligned}
$$

where

S^* is the subset of S containing those records, k, such that such that $w_k > 1$ and $w_k < 6$,

n_S is the size of S*, and

E_k is the inverse of the sampling rate used for k (e.g., if k was sampled at a one-in-six rate, then $E_k = 6$), *which is set to 0 when k is not in the sample*.

An estimate of the mse component from the original nonresponse adjustment is

$$
\begin{aligned}
\mathrm{mse}_N(t_y^S) = & \textstyle\sum^G (1 - [1/a_{(g)}]) \\
& \{[r_g/(r_g - 1)][\textstyle\sum_{k\in Ug}(a_k y_k^S)^2 - (\sum_{k\in Ug} a_k y_k^S)^2/r_g] - \sum_{k\in Ug}[E_k^{\,2} - E_k)w_k^{\,2}e_{kS}^{\,2}\}, \qquad (D7)
\end{aligned}
$$

where $y_k^S = \mathbf{g}_k\mathbf{b}^S + E_k(w_k/a_k)e_{kS}$ is defined for all of U (it is moot that e_{kS} is unknown for k not in the sample, since $E_k = 0$ for such records). The last term is a bit *ad hoc*. It attempts to remove the added noise from using sample values within the nonresponse-variance calculation.

Let S_h (or Sh) denote the subset of all sample records in sample reweighting group h (h = 1, ..., H) after collapsing. An estimator for the mse component of t_y^S due to sampling and integerization is

$$mse_S(t_y^{S(I)}) = \sum_S L_k^{(I)}(L_k^{(I)} - w_k)u_k^2, \qquad \text{(D8)}$$

where

u_k = $[n_{Sh}/(n_{Sh} - 1)]^{1/2} (y_k - R_h^S)$ for $k \in S_h$,

n_{Sh} is the number of operations in S_h, and

R_h^S = $\sum_{Sh} L_k^{(I)}y_k / \sum_{Sh} L_k^{(I)}$.

(Note: In the rare situation where $n_{xh} = 1$, set $u_k = y_k / \surd 2$. This is not unbiased, but it is reasonable, and we need to do something.)

The estimated total mean squared error for a sample number is

$$mse_T(t_y^{S(I)}) = mse_C(t_y^S) + mse_N(t_y^S) + mse_S(t_y^{S(I)}). \qquad \text{(D9)}$$

The square root of this value is the estimated root mean squared error for a sample number, which is displayed on Table B of *Appendix C*. The contribution from coverage adjustment, which in this form excludes integerization, is

$$R_C = mse_C(t_y^S)/mse_T(t_y^{S(I)}), \qquad \text{(D10)}$$

which is displayed on Table B in percentage form, along with the contribution from nonresponse adjustment and sampling, $1 - R_C$.

Hawaii

For Hawaii, mse's were estimated using the simple formula:

$$\text{mse}_H(t_y^{(l)}) = \sum_U a_k^{(l)}(a_k^{(l)} - 1)y_k^2.$$

The theory behind this is that each farm has an independent probability of response approximately equal to $1/a_k$. In addition, $a_k^{(l)}$ is a random variable with mean a_k.

US-level Mean Squared Errors

Although there were correlations in the estimates of state-level targets due to smoothing in area-farm based values, these correlations were ignored when computing US-level mse's. Moreover, biases were treated as components of variances, so that an mse at the US level was estimated as the sum of the corresponding mse's at the state level.

Appendix – A slight improvement on equation (D3$^{(l)}$) is

$$\begin{aligned} mse_C(t_y^{C(l)}) \quad = \quad & var_{NML\_x}(b_0 + \mathbf{z}_T\mathbf{b}_z)^2 + \\ & X_{NML}{}^2\mathbf{b}_z\text{'}\ \mathbf{COV}\ \mathbf{b}_z + X^2(\mathbf{bias}_{(l)}\mathbf{b}_z)^2 \\ & \textstyle\sum_U w_k(w_k - a_k)_{w\geq a}\, e_k{}^2 + \sum_U [\ |w_k{}^{(l)} - w_k| - (w_k{}^{(l)} - w_k)^2]\ e_k{}^2\ . \end{aligned}$$

Unlike equation (D3$^{(l)}$), the above expression doesn't rely on $w_k{}^{(l)2}$ estimating its xpectation, and cannot be negative.

Similarly, a slight improvement on equation (D8) is

$$mse_S(t_y^{S(l)}) = \textstyle\sum_S L_k(L_k - w_k)u_k{}^2 \ + \ \sum_S [\ |L_k{}^{(l)} - L_k| - (L_k{}^{(l)} - L_k)^2]\ u_k{}^2.$$

Neither of these improvements were incorporated into the 2002 calculations.

Determining the *Attributed* and *Corrected* Fractions of Nonreponse and Undercoverage Adjustment

For each farm k in a state, let

$a_k^{(I)}$ be the farm's integerized nonresponse-adjusted (Census) weight
$w_k^{(I)}$ be the farm's integerized coverage-adjusted (Census) weight, and
c_k be the farm's corrected nonresponse-adjusted (Census) weight.

The last value, c_k, is identical to $a_k^{(I)}$ for all Census records except (possibly) screener-nonrespondent/census-respondents. Let q be the number of census forms mailed to screener-nonrespondents in the state divided by the number of those forms returned and valid (including valid out of scopes). For a screener-nonrespondent/census-respondent k in this state, c_k is set equal to q.

The c_k that are greater than the corresponding $a_k^{(I)}$ need to be integerized. To that end, put the records in the state for which $c_j > a_j^{(I)}$ into a data set. Sort them by land in farm and then run the re-integerization routine (see pages A7 - A9) on them with respect to the integerization of the corresponding w_k (the pre-integerized coverage-adjusted weight for j). Call the resulting integerized weight for k, $c_k^{(I)}$. (When $c_j = a_j^{(I)}$, $c_j^{(I)} = c_j$.)

The *corrected* nonresponse adjustment expressed as a fraction of a state coverage-adjusted total, $T = \sum w_k^{(I)}y_k$ (the summation is over all Census records in the state), is computed as

$$P_{C\_NR} = \sum (c_k^{(I)} - 1)y_k / \sum w_k^{(I)}y_k. \tag{E1}$$

The corrected coverage adjustment expressed as a fraction of a state coverage-adjusted total, $T = \sum w_k^{(I)}y_k$, is computed as

$$P_{C\_Cov} = \sum (w_k^{(I)} - c_k^{(I)})y_k / \sum w_k^{(I)}y_k. \tag{E2}$$

The *attributed* nonresponse adjustment expressed as a fraction of a state coverage-adjusted total, $T = \sum w_k^{(I)} y_k$, is computed as

$$P_{A\_NR} = \sum (a_k^{(I)} - 1_k) y_k / \sum w_k^{(I)} y_k. \qquad \text{(E3)}$$

The attributed coverage adjustment expressed as a fraction of a state coverage-adjusted total, $T = \sum w_k^{(I)} y_k$, is computed as

$$P_{A\_Cov} = \sum (w_k^{(I)} - a_k) y_k / \sum w_k^{(I)} y_k. \qquad \text{(E4)}$$

The “attributed” fractions refer to how the missing records are attributed to the total. That is to say, screener-nonrespondent/census-nonrespondents is handled through the coverage adjustment rather than the nonresponse adjustment.

In contrast to this, the “corrected” fractions acknowledge that screener nonrespondents were indeed mailed census forms. Some screener norespondents chose to respond to the Census. Others did not. A correct measure of nonresponse recognizes this.

Fractions can be multiplied by 100 to be converted into percentage form. Corrected numbers (from equations (E1) and (E2)) are computed at the US and state level. US-level corrected numbers are displayed in Table A of http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf, what we have called *Appendix C*. Corrected numbers for each state are displayed in Table A of http://www.nass.usda.gov/census/census02/volume1/$$/$$2appxc.pdf (replacing the “$$” with the appropriate state abbreviation).

Table C displays attributed numbers (from equations (E3) and (E4)) at state level in the US-level Appendix C and the county level in a state Appendix C.

Links

The 2002 Census of Agriculture
http://www.nass.usda.gov/census

"Appendix C" (US)
http://www.nass.usda.gov/census/census02/volume1/us/us2appxc.pdf

"Appendix C" (Nebraska)
http://www.nass.usda.gov/census/census02/volume1/ne/ne2appxc.pdf

Fetter, M.J. and Kott, P.S. (2003). "Developing a Coverage Adjustment Strategy for the 2002 Census of Agriculture"
http://www.fcsm.gov/03papers/fetter_kott.pdf